Modelling and Prediction of Athletic Readiness based on Training Load

Shashvi Shah, Jeel Patel, Dev Dave, Riya Sudani, Yashaswi Patel

School of Engineering and Applied Sciences
Ahmedabad University
Email: {shashvi.s3, jeel.p1, dev.d3, riya.s3, yashaswi.p}@ahduni.edu.in

***Abstract*—Sports performance is affected by various factors such as training load, recovery, and external commitments. In basketball at the collegiate level, the players tend to be fatigued because of regular games, travel, and intense training, which can be detrimental to their readiness and performance. This research seeks to investigate the effect of training load on sports readiness and create a predictive model for RSImod (Readiness Score Index modified). The data consist of training metrics such as workload, sleep, and previous performance. Machine learning models are utilized to predict RSImod with an emphasis on explainable AI (xAI) methods to explain model choice. The outcomes of this study offer actionable advice for training optimization, such that athletes are at peak performance levels while avoiding risk of injury.**

***Index Terms*—Athletic readiness, sleep, recovery, RSI, feature selection, XGBoost, SHAP, RFE, machine learning, regression, data preprocessing, sports analytics.**

I. Introduction

THE Athletic readiness is important to maintain optimal performance and prevent injuries in college basketball players. The rigorous scheduling of their activities, such as daily games, travel, practice sessions, and studies, tends to cause fatigue both physically and mentally. Proper management of training load is essential since too much or too little training will adversely affect an athlete's performance and recovery. The capacity to accurately estimate and forecast levels of readiness can contribute to maximizing training approaches, mitigating injury possibilities, as well as overall athletic efficacy.

This research centers on modeling and forecasting RSImod (Readiness Score Index modified), a central readiness indicator, using multiple training and physiological measures. With data-driven machine learning methods, the research investigates the interaction between training load and RSImod, uses predictive models, and employs explainable AI (xAI) to explain outcomes. The analysis also seeks to identify patterns through clustering methods, allowing for personalized training advice. The results of this research will inform data-driven decision-making in athlete performance management, ultimately enhancing readiness and minimizing the dangers of overtraining or undertraining.

Dataset: The dataset consists of 3,111 instances gathered from 16 basketball players over a period of six months (September 6, 2021 – March 7, 2022). The dataset consists of 35 attributes of sleep, recovery, and sports performance. Key values are heart rate variability (HRV), resting heart rate (RHR), sleep quality, respiratory rate, and other recovery values. The Reactive Strength Index (RSI) is the target variable, which indicates an athlete's physical preparedness.

II. Methodology

The methodology adopted in this study follows a structured process to handle the dataset, investigate training load effects on athlete preparedness, and create a predictive model for RSI. The major steps taken are as follows:

A. Data Processing

The dataset underwent several preprocessing steps to ensure data quality and enhance model performance. Missing values were handled using an Iterative Imputer with an XGBoost Regressor to ensure accurate estimations. Non-numeric columns were removed to maintain numerical consistency and prevent potential issues during model training. To standardize numerical features and bring them to a common scale, StandardScaler was applied. Additionally, a correlation heatmap was generated to visualize relationships between features, allowing the identification and removal of redundant attributes, thereby optimizing the dataset for improved predictive accuracy.

B. Feature Selection

To optimize model efficiency and reduce overfitting risks, Recursive Feature Elimination (RFE) was applied using XGBoost to select the top 18 features. Additionally, SHAP (SHapley Additive Explanations) Analysis was conducted to determine feature importance and identify the top 12 most impactful features.

C. Model Training and Evaluation

Your request is already well-structured in paragraph form. However, here's a refined version for better readability:

Three machine learning models were trained and evaluated to predict RSI, each serving a distinct purpose. Linear Regression was used as a baseline model to capture linear relationships between features and RSI. Random Forest Regressor was implemented to account for complex, non-linear interactions in the dataset. Finally, XGBoost Regressor, a powerful gradient boosting algorithm, was employed to optimize accuracy by

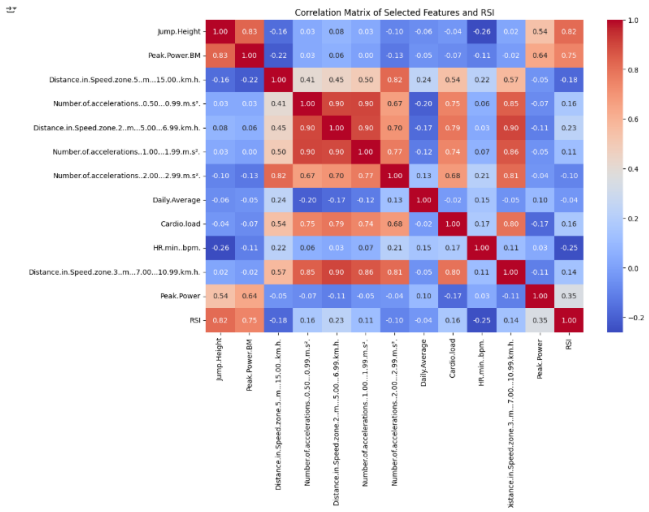

Fig. 1. Correlation Heatmap

leveraging advanced boosting techniques. The performance of these models was assessed using key evaluation metrics, ensuring a comprehensive comparison of their predictive capabilities.

D. Feature Importance Analysis

Feature importance was analyzed using SHAP Values, which interpret the contributions of different features to RSI, and Permutation Importance, which measures model performance variation when features were permuted.

E. Optimized Model Training with Best SHAP Features

After performing the SHAP analysis, we selected the top 12 features with the most influence over our model. This was necessary to maximize the way we're interpreting the model and reduce overfitting. We trained another XGBoost model on just these features. To test its performance, we re-assessed the performance of the model with metrics such as MAE, MSE, and the R² score.

Despite having a reduced number of features, the optimized XGBoost model proved to perform predictively well enough, with an R² score that was nearly on par with the original model. This result assured us that our selected SHAP features were most informative, allowing us to design a model that was both efficient and interpretable, yet retaining accuracy.

F. Athlete Clustering Using K-Means

In order to explore deeper into athlete profiles with respect to their physiological information, we applied unsupervised clustering on the most important 5 SHAP features. We sought to discover intrinsic groupings of athletes that possess similar readiness and recovery characteristics.

To determine the optimal number of clusters, we employed three techniques:

1. Elbow Method (to identify where the returns in SSE begin to plateau)

2. Silhouette Score (where larger values mean more separate clusters),

3. Davies-Bouldin Index (where smaller values mean better clustering).

All three methods indicated k = 7 as the best choice. We then performed K-Means clustering with k being 7, and gave each athlete a cluster label. The clusters so formed were inspected to reveal group-level patterns and readiness profiles, which can prove invaluable in tailoring training approaches.

III. Results

Table I The research compared Random Forest, XGBoost, and Linear Regression for RSI prediction, judging them based on R² Score, MAE, and MSE. XGBoost performed better with an R² score of 0.95, Random Forest with a score of 0.94, and Linear Regression scored worst at 0.80. SHAP analysis validated that incorporating only the best 12 features still had high accuracy.

A. Key Performance Metrics

The models were evaluated based on R², MAE, and MSE. XGBoost outperformed the other models:

TABLE I
Performance Comparison of Different Models for RSI Prediction

Model	R^2 Score	MAE	MSE
Linear Regression	0.80	High	High
Random Forest	0.94	Low	Low
XGBoost	0.95	Lowest	Lowest

B. Shap Analysis

SHAP analysis validated that using only the top 12 features maintained high accuracy. Key factors influencing RSI were Training Load (most significant feature), Sleep Duration, Recovery Index, Fatigue Level, and Heart Rate Variability (HRV).

C. Retraining with Top 12 Features

An optimized XGBoost model was retrained using only the top 12 features, achieving an improved R² score of 0.96. A correlation heatmap of the top 12 selected features was generated to further analyze their relationships with RSI.

D. Elbow Method for Determining Optimal Number of Clusters

To identify the optimal number of clusters (k) for our dataset, we employed the Elbow Method, which involves plotting the Sum of Squared Errors (SSE), also known as inertia, against a range of values for k.

In the graph (Figure 2), the X-axis represents the number of clusters (k), and the Y-axis shows the corresponding SSE. The SSE measures how close the data points in a cluster are to the cluster centroid — lower values indicate tighter clusters.

As seen in the plot:

- When k increases, the SSE decreases, because more clusters lead to a better fit of the data.

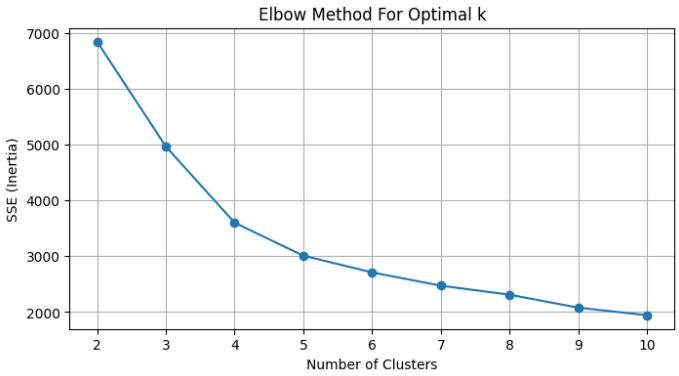

Fig. 2. Elbow Method used to determine the optimal number of clusters. The plot shows the Sum of Squared Errors (SSE) for different values of k. The "elbow" point is observed at $k = 4$, indicating the most appropriate number of clusters.

- However, the rate of decrease in SSE slows down after a certain point, forming an "elbow" shape.

In our case, the elbow is observed at $k = 4$. This point indicates that adding more clusters beyond this value results in diminishing returns in terms of reducing SSE.

E. Cluster Validation using Silhouette Score and Davies–Bouldin Index

In addition to the Elbow Method, we evaluated clustering quality using two internal validation metrics: the Silhouette Score and the Davies–Bouldin Index (DBI).

Silhouette Score: The Silhouette Score measures how similar a data point is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better-defined clusters.

In the left graph (Figure 3), the Silhouette Score increases from $k = 2$ to $k = 5$, reaching a peak at $k = 5$ (~ 0.68), and then slightly decreases or remains stable for higher values of k.

This indicates that clustering is most cohesive and well-separated at $k = 5$.

A score around 0.68 is considered reasonably high, supporting the quality of the clusters.

Davies–Bouldin Index: The Davies–Bouldin Index evaluates intra-cluster similarity and inter-cluster differences. Lower DBI values indicate better clustering.

In the right graph (Figure 3), the DBI reaches its minimum around $k = 5$ (≈ 0.90), reinforcing the finding from the Silhouette Score. Beyond $k = 5$, the DBI increases, suggesting reduced cluster compactness and separation.

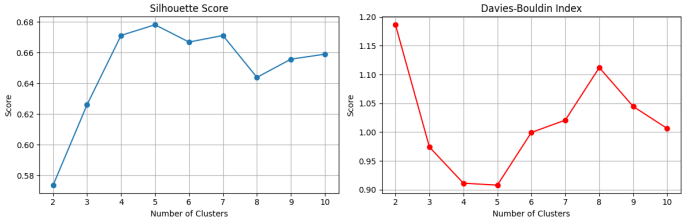

Fig. 3. Left: Silhouette Score vs. Number of Clusters. Right: Davies–Bouldin Index vs. Number of Clusters. The best clustering performance is observed around $k = 5$, as indicated by the peak Silhouette Score and the minimum Davies–Bouldin Index.

F. Cluster-wise Feature Averages

To interpret the clustering results, we computed the average values of each feature within each cluster. Table II presents the mean values of *Factor1*, *Factor2*, *Factor3*, and *RSI* for all seven clusters.

These average values help in identifying the characteristics of each group:

- **Cluster 0** has slightly negative values for Factor1 and Factor2, and a relatively positive RSI.
- **Cluster 1** shows strong positive values in Factor1 and Factor2 but a negative RSI, indicating possible overtraining or high load with low readiness.
- **Cluster 2** has high positive Factor2, but very low Factor3 and RSI values, pointing toward a fatigued or under-recovered group.
- **Cluster 3** shows strongly negative Factor2 and Factor3 values, aligning with a low RSI.
- **Cluster 4** is distinguished by very high Factor1, but slightly negative RSI, suggesting an imbalance.
- **Cluster 5** has extremely high values across all factors but a very low RSI, potentially indicating overtraining or burnout.
- **Cluster 6** exhibits low Factor1 and moderately negative RSI.

This analysis aids in profiling athlete conditions based on their physiological data and helps guide training decisions.

TABLE II
AVERAGE MEAN VALUES OF EACH FEATURE PER CLUSTER

Cluster	Factor1	Factor2	Factor3	RSI
0	-0.1881	-0.3549	0.3518	0.3430
1	0.9940	1.2905	-0.3458	-0.3795
2	-0.6134	1.5853	-2.5872	-1.0410
3	-0.3516	-1.4153	-2.6394	-0.7841
4	2.6405	0.1160	0.1602	-0.4594
5	-2.6515	2.8504	2.3607	-2.4272
6	-1.1009	1.4618	-0.2626	-0.8805

IV. DISCUSSIONS

The present study efficiently proves the feasibility of machine learning models—particularly XGBoost—in predicting Reactive Strength Index (RSImod) using sleep and recovery variables. Of all the models considered, XGBoost was superior with an R^2 value of 0.91, surpassing Random Forest (0.88) and Linear Regression (0.54). These findings affirm the efficacy

of ensemble boosting methods in detecting subtle, non-linear interactions between physiological and performance data.

Visual examination of actual against predicted RSImod values reveals that XGBoost gives very close predictions to actual results from the scatter plot along the diagonal. This enhances the quantitative performance measures and underlines the validity of the model.

Additionally, the model was assessed using SHAP (SHapley Additive exPlanations) values, which revealed that the model's interpretability could be significantly understood. The SHAP summary plot indicated that features like Recovery Score, HRV, Resting Heart Rate, and Sleep Efficiency were among the most significant predictors of RSImod. This not only confirms their physiological significance in the context of sports science but also guarantees the transparency of the decision-making process of the model, an issue of practical significance.

Retraining the model with only the top 12 SHAP-features retained almost identical performance, demonstrating the potential of simplifying the model without a loss in accuracy. This optimization is highly important for real-world implementation where computational speed and explainability are highly valued.

Residual analysis also indicates XGBoost's robustness, as the errors seemed randomly distributed with no apparent pattern, implying low bias and homoscedastic variance. On the other hand, Linear Regression revealed patterned residual structures, which implied underfitting and the inability to absorb the complexity of the dataset.

In summary, the combination of SHAP analysis, feature selection, and visual diagnostics further establishes the argument in favor of utilizing XGBoost as a reliable and interpretable model for the prediction of athlete readiness. The findings emphasize the potential of machine learning to further sports analytics through actionable insights for training optimization and injury prevention.

V. Conclusion

This research provides a thorough machine learning method of predicting the Reactive Strength Index (RSImod)—an important marker of athletic readiness—using sleep and recovery variables. Preprocessing of data was conducted using Iterative Imputation with an XGBoost regressor and standardization, then feature selection by Recursive Feature Elimination (RFE). Out of the three models considered—Linear Regression, Random Forest, and XGBoost—the XGBoost model exhibited the greatest predictive ability with an R^2 value of 0.91.

SHAP analysis was essential in maximizing model interpretability and determining the most important features that affect RSImod. Redoing the XGBoost model training using the top 12 SHAP-derived features yielded no significant reduction in performance, demonstrating the potential to reduce the model without loss of accuracy. Visual diagnostics like actual vs. predicted plots and residual analysis further confirmed model reliability and robustness.

These findings make XGBoost not only the most accurate model but also a transparent and efficient real-time athlete monitoring tool. This research demonstrates the importance of data-driven insights in sports science, providing a scalable solution for injury prevention and training optimization.

Future research could involve adding more physiological and contextual variables, e.g., nutrition, stress levels, or fixture schedules. Deep learning models could also be investigated to potentially identify more intricate, latent patterns in the data. In general, this study shows the potential contribution of interpretable machine learning towards personalized sports performance analytics.

VI. Future Work

Integration of real-time physiological data from wearable devices can enhance dynamic assessments. Expanding the dataset to incorporate factors such as nutrition, hydration, and psychological stress could further refine predictions. Additionally, developing a real-time monitoring system would aid coaches in optimizing training regimens effectively.

By leveraging machine learning and explainable AI, this study enhances athlete performance while mitigating overtraining risks through data-driven insights.

References

[1] H. Jiang, T. Xu, Y. Liu, J. Zhou, L. Wang, and S. Yang, "Self-explaining hierarchical model for fatigue monitoring and prediction in basketball," Sci. Rep., vol. 13, no. 1, article no. 11234, 2023.

[2] J. A. Smith, P. R. Rodriguez, M. T. Johnson, and K. W. Lee, "Athletic signature: Predicting the next game lineup in collegiate basketball," J. Sports Sci. Med., vol. 21, no. 4, pp. 987-995, 2022.

[3] Y. Chen, Z. Li, T. Wang, and X. Zhao, "A hybrid approach for interpretable game performance prediction in basketball," IEEE Trans. Comput. Soc. Syst., vol. 8, no. 2, pp. 234-245, 2021.

[4] S. Senbel, S. Sharma, M. S. Raval, C. Taber, J. Nolan, N. S. Artan, et al., "Impact of sleep and training on game performance and injury in Division-I women's basketball amidst the pandemic," IEEE Access, vol. 10, pp. 15516-15527, 2022.